REMARKS

In response to the Office Action mailed February 22, 2010, the Assignee of the present application (*Nuance Communications, Inc.*) respectfully requests reconsideration. Claims 1, 4-5, 7-11, 14-15, 17-23 and 29-35 were previously pending for examination. Claims 1, 7-9, 11, 17-19 and 21 have been amended herein. No claims have been canceled or added. As a result, claims 1, 4-5, 7-11, 14-15, 17-23 and 29-35 remain pending for examination, with claims 1, 11 and 21 being independent. No new matter has been added.

The claim amendments are supported throughout the specification. For example, support for the amendments to independent claims 1, 11 and 21 can be found in the specification at least at page 15, line 24 – page 16, line 5; and page 18, lines 9-11.

Rejections Under 35 U.S.C. 103

The Office Action rejects claims 1, 4, 7-8, 10-11, 14, 17-18, 20-23 and 33-35 under 35 U.S.C. 103(a) as purportedly being obvious over Lumelsky (U.S. Patent No. 6,081,780) in view of Karaali (U.S. Patent No. 5,668,926). The Office Action rejects claims 5, 9, 15, 19 and 29-32 under 35 U.S.C. 103(a) as purportedly being obvious over Lumelsky in view of Karaali and one or more other references. The Assignee respectfully traverses these rejections.

I. Overview of Some Embodiments

The present application describes systems and methods for converting a text input to synthesized speech in a manner that mimics the style and pronunciation of a spoken example of the text input (page 1, lines 6-10). When a user supplies a spoken example of a text string to be text-to-speech synthesized, some embodiments can extract prosodic parameters from the spoken example, and adopt those prosodic parameters as synthesis parameters for generating the synthetic speech waveform (page 8, lines 1-8). Prosodic parameters are speech quality attributes, such as pitch,

duration, energy values, etc. for multiple speech segments in the spoken audio signal (page 1, lines 21-42; page 10, lines 23-24).

Some embodiments allow a user to input a text string and speak an example of a desired pronunciation of the text string (page 9, lines 9-12). A prosody analyzer can then process the spoken audio signal to extract prosodic parameters from it, e.g., pitch and energy contours (i.e., sets of pitch and energy values extracted from various time indexes during the audio signal) (page 10, line 20 – page 11, line 1). An alignment module can extract the duration of each unit of the text (e.g., word or phoneme) as produced in the spoken example (page 13, lines 18-21). This can be accomplished by aligning the audio signal with the text.

After prosodic and durational parameters have been extracted by the prosody analyzer and adopted as synthesis parameters, a conversion module can process the synthesis parameters to generate an input for a text-to-speech (TTS) engine (page 15, line 24 – page 16, line 4). The input may be in a format such as SSML (Speech Synthesis Markup Language), which provides the text string with prosody markup that can be processed by the TTS engine to produce a synthetic speech waveform with the specified prosodic parameters (page 16, line 9 – page 17, line 7). The duration parameters extracted by the alignment module allow the prosodic parameters from various time indexes during the audio signal to be mapped to the appropriate text units when the text is converted to synthetic speech (page 16, lines 5-8). The TTS input so generated may then be used by the TTS engine in converting the text string to synthetic speech (page 18, lines 9-11).

The foregoing overview is provided to assist the Examiner in appreciating some aspects of the invention. However, this overview may not apply to each of the independent claims, and the language of each independent claim may differ in material respects from the overview above. Therefore, the Assignee respectfully requests that careful consideration be given to the language of each independent claim, and that each be addressed on its own merits, without relying on the overview above. In this respect, the Assignee does not rely on the overview above to distinguish any of the claims over the prior art, but rather relies only upon the claim language and the arguments presented below.

II. Overview of Lumelsky

Lumelsky describes adjusting a phonetic representation of a text used in synthesizing that text to speech. The adjustment is made through corrective feedback based on spectral comparison of an audio signal produced by the TTS system and an audio signal spoken by a human narrator (Lumelsky: col. 9, lines 1-6). When the system receives a text to be synthesized, the system generates initial prosodic parameters with reference to an internal dictionary, rather than with reference to a spoken example (Lumelsky: col. 12, line 44 – col. 13, line 8). A first synthetic speech audio signal is then generated using the initial prosodic parameters, and spectrally compared to a spoken audio signal from the narrator (Lumelsky: col. 13, lines 42-58). The spectral distance between the synthetic audio signal and the narrator's audio signal then determines the amount of correction to be applied to the TTS system's prosodic parameters in the next synthesis iteration (Lumelsky: col. 14, lines 17-25). Synthesis and correction iterations are repeated until satisfactory results are achieved (Lumelsky: col. 14, lines 25-27).

III. Overview of Karaali

Karaali describes a system that converts text to speech (Karaali: Abstract). Text that is to be converted is first translated into a series of phonetic representations (Karaali: col. 2, lines 66-67). A duration is then assigned to each phonetic representation by a rule-based component (Karaali: col. 3, lines 34-40). An acoustic representation is assigned to each frame of each phonetic representation by a neural network, and a synthesizer then converts the series of acoustic representations to an audio signal (Karaali: col. 17, lines 57-58).

Before the neural network can be used as described above in the process of converting text to speech, it must be trained on some training data (Karaali: Abstract). The training data consists of a training text and a spoken recording of the training text (Karaali: col. 18, lines 24-28). The training text is transcribed to a phonetic form which is time aligned with the recorded audio to map a varying number of audio frames to each phone represented in the text (Karaali: col. 18, lines 28-40). The neural network is then trained to associate each phonetic representation frame from the

training text with the acoustic representation of the corresponding audio frame from the spoken recording (Karaali: col. 18, line 60 – col. 19, line 30).

IV. Independent Claim 1 Patentably Distinguishes Over Lumelsky and Karaali

Independent claim 1 as amended recites, *inter alia*, "adopting as synthesis parameter values the prosodic parameter values and duration parameter values extracted from the audio signal; and generating a synthetic speech waveform using the synthesis parameter values." Neither Lumelsky nor Karaali discloses or suggests these limitations.

As discussed above, Lumelsky adjusts parameters for text-to-speech synthesis through an iterative process. In the first iteration, a first synthetic speech waveform is generated from a text using initial prosodic parameters that are generated using an internal dictionary, without reference to the spoken audio signal. It is only after an initial synthetic speech waveform is generated that the spoken audio signal is considered, and even then the synthesis parameters for the next iteration are only adjusted based on a comparison between the spoken audio and the previously generated synthetic audio. Lumelsky never adopts parameter values extracted from the spoken audio as synthesis parameter values, as required by claim 1. The only synthesis parameters that Lumelsky determines are taken from a dictionary, and are not specified as any parameter values extracted from an audio signal. These parameters are then adjusted based on a distance between audio signals, without ever adopting parameter values extracted from an audio signal as synthesis parameter values.

Karaali does not cure the deficiencies of Lumelsky in this respect. As discussed above, Karaali's system only makes use of a spoken audio recording in training a neural network to associate acoustic representations with phonetic representations, not in the actual process of synthesizing text to speech. When Karaali's system converts a text to speech, no audio signal corresponding to a pronunciation of that text is involved. Karaali does not adopt any parameter values extracted from an audio signal as synthesis parameter values for use in generating a synthetic speech waveform.

Even if Lumelsky and Karaali were combined, the alleged combination would fail to meet at least the above-discussed limitations of claim 1. Therefore, claim 1 patentably distinguishes over any combination of Lumelsky and Karaali, and it is respectfully requested that the rejection of claim 1 be withdrawn.

Claims 4-5, 7-10, 29-30 and 33 depend from claim 1 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims be withdrawn.

V. Independent Claim 11 Patentably Distinguishes Over Lumelsky and Karaali

Independent claim 11 as amended recites, *inter alia*, "adopting as synthesis parameter values the prosodic parameter values and duration parameter values extracted from the audio signal; and generating a synthetic speech waveform using the synthesis parameter values."

For reasons that should be clear from the foregoing discussion of Lumelsky and Karaali, these references, whether alone or in combination, fail to disclose or suggest at least the above limitations of claim 11. Therefore, claim 11 patentably distinguishes over any combination of Lumelsky and Karaali, and it is respectfully requested that the rejection of claim 11 be withdrawn.

Claims 14-15, 17-20, 31-32 and 34 depend from claim 11 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims be withdrawn.

VI. Independent Claim 21 Patentably Distinguishes over Lumelsky and Karaali

Independent claim 21 as amended recites, *inter alia*, "a conversion module for adopting as synthesis parameter values the prosodic parameter values and duration parameter values extracted from the audio signal; and a TTS engine for generating a synthetic speech waveform using the synthesis parameter values."

For reasons that should be clear from the foregoing discussion of Lumelsky and Karaali, these references, whether alone or in combination, fail to disclose or suggest at least the above limitations of claim 21. Therefore, claim 21 patentably distinguishes over any combination of Lumelsky and Karaali, and it is respectfully requested that the rejection of claim 21 be withdrawn.

Claims 22-23 and 35 depend from claim 21 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims be withdrawn.

General Comments on Dependent Claims

Because each of the dependent claims depends from a base claim that is believed to be in condition for allowance, the Assignee believes that it is unnecessary at this time to argue the further distinguishing features of all of the dependent claims. However, the Assignee does not necessarily concur with the interpretation of the dependent claims as set forth in the Office Action, nor does the Assignee concur that the basis for the rejection of any of the dependent claims is proper. Therefore, the Assignee reserves the right to specifically address in the future the further patentability of the dependent claims not specifically addressed herein.

CONCLUSION

In view of the foregoing, the present application is believed to be in condition for allowance. A Notice of Allowance is respectfully requested. The Examiner is requested to call the undersigned at the telephone number listed below if this communication does not place the case in condition for allowance to discuss any outstanding issues relating to the allowability of this application.

If the response is not considered timely filed and if a request for an extension of time is otherwise absent, the Assignee hereby requests any necessary extension of time. The Assignee believes no fee is due with this response. However, if a fee is due, please charge Deposit Account No. 23/2825 under Docket No. N0484.70760US00 from which the undersigned is authorized to draw.

Dated: 5/24/10

Respectfully submitted,
Nuance Communications, Inc.

By _____
Richard F. Giunta
Registration No.: 36,149
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, Massachusetts 02210-2206
617.646.8000